Insights on road safety with open data: the case of Rome

Antonio Miloso[1,*], *Eleonora* Veglianti[2], *Marco* De Marco[3], and *Elisabetta* Magnaghi[2]

[1]Expertlab s.r.l., Rome, Italy.
[2]FGES- Université Catholique de Lille, 59800 Lille, France.
[3]Department of Economics, University of Uninettuno, 00186 Rome, Italy.

Abstract. Modern cities face the challenge of providing citizens with an appropriate level of services to maintain the growing population. Thanks to the support of open data policymakers are capable of ensuring administrative transparency and participation in decisions, enabling citizens and employees to effectively use services and tools and integrating physical and intangible infrastructures (systems, data and processes) in a service-oriented perspective. This study investigates open data about car accidents in the metropolitan city of Rome between 2014 and 2019 through the service science lens. It is pointed out how the city roads maintenance (for example, road surfaces, road signs and traffic extent) can significantly affect the number of people involved in accidents. From these results, possible improvements in diminishing the number of people involved in car accidents are explored through a prescriptive analysis. This study represents a powerful tool to improve services in the public sphere and an example of the shared value generated by open data initiatives. It contributes in improving the understanding of a data-oriented culture and of building a network of people in all public administrations to increase the shareable information assets of the metropolitan city of Rome.

1 Introduction

The urban populations and the continuous economic growth in urban areas require a new agenda for governments to deal with a more efficient urban planning which has an impact not only in terms of accidents but for the overall sustainability of public services.

Therefore, there is much to gain from doing so from a social as well as from an economic perspective, considering the overall accidents and its victims. On one side, there are human and health care costs (i.e. medical treatment). On the other side, there are costs related to accident itself and to the administrative issues such as damages to vehicles, legal conflicts and cost of intervention.

In this context, local administrators have to do an accurate urban planning process and analyses to identify the main factors to reach the zero-accident goal given by the EU within 2050. Open data platforms represent a boundary-redefining technology [1]–[3] with a new

[*] Corresponding author: antonio.miloso@expertlab.it

setting that is changing data and information linkages. Therefore, open data implementation presents a dilemma for public managers, especially to risk-averse individuals [4], [5].

This paper tackles an important issue taking place worldwide, in general, and in Italy. In particular, it shows how it is possible to implement useful strategies and valuable insights into the public sector with open data. From the service science perspective, this article explores open data platforms as an opportunity to improve the relationship between the public sector and citizens/users, to decrease costs providing new and more efficient services.

With this promise, this paper wants to contribute to the existing literature about service science studying the role of open data in improving road safety. Consequently, the research question is the following: which are the roads features that have a relevant impact on accident extents?

This paper is structured as follows. Section 2 presents the literature review; Section 3 describes the research methodology and adopted approach; Section 4 presents the results. Finally, Section 5 concludes, highlighting the findings, analysing the limits of the paper and suggesting further works.

2 Literature review

The present growth of the service sector in global economies is unique in human history [6]. The need for service innovations to enhance additional economic growth is crucial in different contexts.

Service science is defined as an integration of various disciplines such as management, engineering, accounting, finance and operations, to use the next set of innovators to contribute to the service economy [7]. Service systems are dynamic configurations of several elements (i.e. people, technologies, organizations and information) that create and deliver value to different stakeholders [7]. Nowadays, the main challenge is to become more systematic about innovating in service.

Innovation implies the birth of new ideas and/or methods that help to put the new idea into practice. It represents a combination of creativity and implementation [8]. In the service field, innovation brings changes to a service system influencing its evolution [6], [9].

In other words, innovation represents any "idea, practice, or object that is perceived as new by an individual or other unit of adoption" [10], [11]. Therefore, the subsequent adoption of ideas, practices and objects by the same organization is not an innovation anymore [12]. In addition, innovations are a break from the previous setting being different from the incremental change [13]. These elements are evident in open data platforms. Indeed, the organizational decision of adopting open data represents a crucial change from prior transparency policies and ICT systems since raw data are more available than actual information.

Indeed, often open data platforms contain raw data more than actual information. Information consists in data with a structure and meaning. The process of transforming data into information requires several choices, such as data processing and model selection to extract information.

In this scenario, who extract the information, can be a public employee and the information extracted is influenced by institutional knowledge, expertise and views [14], [15]. However, when data is downloaded and interpreted by an outsider, the organization loses control over the data and, thus, the possible meaning and the extracted information.

In line with this, open data used by people who are not part of the public sector can generate information allowing the information itself to reach the public systems in a direct way providing new useful insights. This feature reveals a critical change in the service science in terms of the public sector and, specifically, in the nature of the interaction between the government sphere and the public [16]–[18].

Open data presents an innovative potential that is not limited to the public sector replacing prior processes in various organizations toward an innovative scenario [19], [20].

In other words, open data bring a revolution in the public service field because it deletes historically bureaucratic behaviours characterized by complex hierarchical structures and peculiar information development [21]. The perspective of control over information in the public service shows direct consequences for the design of public sector ICT systems [22], [23]. For this reason, ICT represents a political concern for public managers [24].

As the relevant literature suggests the “digital-era governance” (DEG) replaced the New Public Management (NPM) to promote government efficiency in the overall service [25]. Following this new conceptual framework, there is a reshaping of prior relationships between the public sector and technology levering open technology to digitize services efficiently [26], [27]. Therefore, service digitization changes the citizen-institutional interaction because service users are able to access directly with core public systems free from institutional-administrative constraints [28], [29].

Implementing open data platforms implies significant consequences. As some scholars suggested, open data improve transparency and public trust [30]–[34]. Others argue that with more data available it is necessary to provide additional information that can increase the efficiency of the government and promote economic growth [35]–[37]. At the same time, some drawbacks emerge as, for instance, the misuse of data [38], [39] and the abuse which can decrease citizen’s trust toward the public service [4], [40]. Moreover, open data can be a security threat [41], [42]. In other words, innovations as open data need an organizational transformation that can be perceived risky in the public service sphere [5], [12], [43], [44]. However, the decision to innovate is driven by the goal of improving the effectiveness of the service [19], [45], [46].

3 Methodology

The adopted methodology applies a machine-learning algorithm to extract the most important features related to the dependent variable [47]. Then, those features are investigated through ANOVA and multiple linear regression to evaluate their relationship with the dependent variables. To accomplish the aforementioned tasks the data are retrieved and pre-processed before sampling to obtain a suitable set of data to train the Gradient Boosting Classifier [48], [49].

3.1 Data gathering

The data used for this study were gathered thanks to an initiative promoted by “Roma Capitale” as part of the open data initiative born according to the National guidelines for the enhancement of public information assets [50]. The purpose of the initiative is to empower citizens to reuse and integrate the data made available to them, to develop services and applications for the benefit of the entire community of users.

For the purpose of this study, only a sample of the data available was extracted. The data used for the analysis describe a sample of 174,222 accidents registered in the metropolitan area of Rome (including the location of Ostia) between the 1st January 2014 and the 11th December 2019. The dataset contains the list of road accidents that occurred in the territory of Roma Capitale in 2014. The dataset includes all the road accidents in which a patrol of any Group of the “Roma Capitale” Local Police intervened. Therefore, accidents in which the parties involved have reached a conciliation are excluded. The dataset does not include the accidents that occurred on the "Grande Raccordo Anulare", is a ring-shaped orbital motorway that encircles Rome.

3.2 Data preparation

The data were accordingly transformed, removing all the unnecessary variables or the ones with a considerable presence of missing values. The accidents reported during the period of the snowing days in 2018 were removed since they described a statistical anomaly.

The resulting dataset contained the following independent variables:

- Road Characteristics;
- Road Type;
- Road Surface;
- Paving;
- Road Signs;
- Atmospheric condition;
- Traffic;
- Visibility;
- Road Illumination.

A dependent variable was computed as the number of human casualties involved in the accident with the sum of the number of people injured plus the number of dead people plus the number of people with reserved prognosis. Then, for classification purpose, another dependent binary variable was generated with a value equal to zero, if there were no human casualties recorded for that specific accident, and equal to one, if there were at least one human casualty.

In the end, the categorical data were transformed into dummy variables to perform the analysis and the observations are sampled (n=140,250) to obtain a homogeneous distribution of the dependent variable.

3.3 Modelling

First, we calculated some descriptive statistics about the distribution of the variables to show the basic characteristics of the recorded accidents. Then the machine learning algorithm Gradient Boosting Classifier was applied to extract the relevant features that defined the classification between accident without and with human casualties.

Gradient boosting is a generalization of boosting to arbitrary differentiable loss functions. This algorithm is an accurate and effective procedure that can be used for both regression and classification applications [49], [51].

We applied the algorithm to a binary classification problem. In according to this algorithm, regression trees are fitted on the negative gradient of the binomial deviance loss function. Since we are using the Gradient Boosting Classifier for a binary classification only a single regression tree is used in this case [48].

The decision trees created during the algorithm processing were analysed to understand what are the important features and how they classify the dependent variable. Individual decision trees already perform feature selection by selecting appropriate split points, which can be used to measure the importance of each feature depending on how often a feature is used to split the tree. More a feature is important, more will cause the tree to split.

Once the most relevant features are extracted their relation to the dependent variable is investigated furthers.

4 Results

4.1 Feature Importance interpretation with gradient boosting classifier

Once the dataset has been appropriately transformed was possible to apply the classification algorithm to interpret the feature importance. The Machine Learning Algorithm adopted the Gradient Boosting Classifier and was run on a dataset of 140,250 observations. The 80% of them was used for the training set and the remaining part (20%) for the test set.

We used the binomial deviance loss function; the learning rate was equal to 0.102 and the number of estimators equal to 250. The Gradient Boosting Classifier was Cross Validated with k-fold equal to 10. The chosen configuration seems accurate, as it is possible to see from the learning curve, shown in Figure 1. The resulting accuracy was 57%.

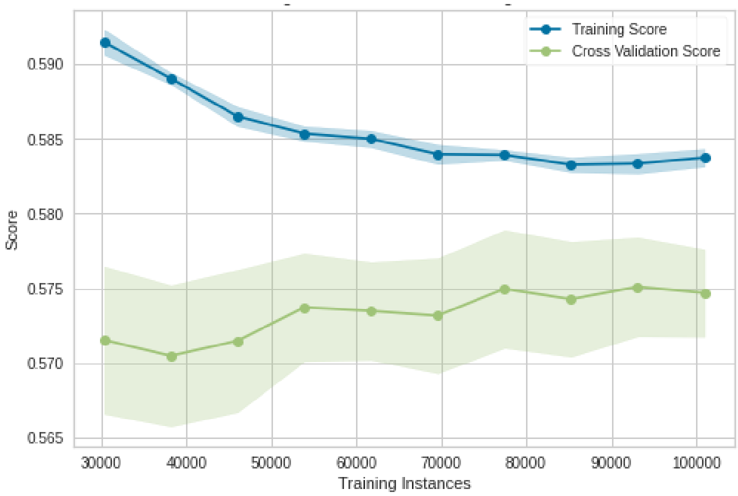

Fig. 1. Learning Curve for Gradient Boosting Classifier.

One of the objectives of this empirical research is to contribute to the knowledge about car accidents in Rome. In doing so, the aim is to show how is it possible to extract useful insights for the public sector and the citizens. The algorithm helped us to identify the determinant factor that concurs in the estimation of a bad accident. As it is possible to see in Figure 2, the most relevant feature is the presence of road holes on the streets.

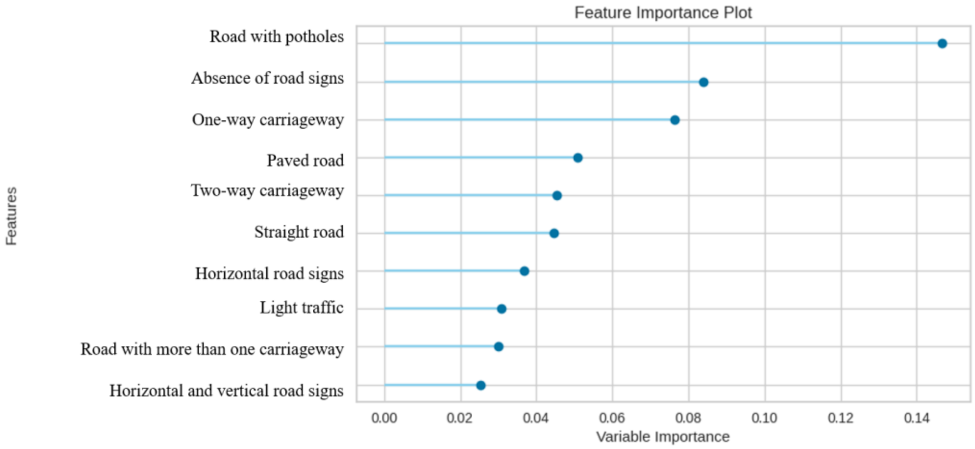

Fig. 2. Feature importance plot.

4.2 Features analysis

The ten features extracted were interpreted with an ANOVA test to investigate the presence of a significant relationship with the dependent variable. Furthermore, their relation was investigated with a multiple linear regression on the number of casualties to understand how those features influence the number of human casualties during the accidents.

An n-way ANOVA was run on a sample of 174,222 accidents to examine the effect of the features (identified by the Gradient Boosting Classifier) on the severity of an accident. All the features has a significant interaction with the dependent variable about the severity of an accident (F(10, 174211) = 431.62, p < 0.001):

- Road with potholes (F(1, 174211) = 593.70, p < 0.0001)
- Absence of road signs (F(1, 174211) = 64.93, p < 0.0001)
- One-way carriageway (F(1, 174211) = 278.68, p < 0.0001)
- Paved road (F(1, 174211) = 6.55, p = 0.0105)
- Two-way carriageway (F(1, 174211) = 164.85, p < 0.0001)
- Straight road (F(1, 174211) = 298.13, p < 0.0001)
- Horizontal road signs (F(1, 174211) = 198.20, p < 0.0001)
- Light traffic (F(1, 174211) = 103.34, p < 0.0001)
- Road with more than one carriageway (F(10, 174211) = 138.28, p < 0.0001)
- Horizontal and vertical road signs (F(10, 174211) = 176.78, p < 0.0001)

Multiple linear regression was run to predict the number of human casualties from the ten variables identified. These variables are statistically significant, F(10, 174211) = 431.62, p < .0001.

From the further inspection of the individual regression, it is possible to observe that the presence of potholes on the streets has a negative effect on the number of human casualties (T = -24.37, p < .0001). Potholes on the streets of Rome are a common topic often abused by the local press. From the analysis, it seems reasonable to assert that in presence of potholes on the road the drivers are slightly more careful since is a notorious cause of accident and damage to vehicles.

No road signs on the street negatively influences the number of human casualties (T = -8.06, p < .0001). The fact that there are no road signs in sight could be related to a less dangerous street than others. In the same way, streets that need more signs are the ones notoriously more dangerous. In effect, the presence of horizontal road signs (T = 14.08, p < .0001) and horizontal and vertical road signs together (T = 13.30, p < .0001) are both positively related with the number of human casualties.

On a one-way carriageway the relation with the number of human casualties in an accident is negative as well (T = -16.69, p < .0001). However, the relation with a two-way carriageway is positive (T = 12.84, p < .0001) as for roads with more than one carriageway (T = 11.76, p < .0001). It seems that wider roads concur in the possibility of higher number of human casualties involved since a car out of control has more space to skid.

Light traffic has a negative effect on the dependent variable (T = -10.17, p < .0001) perhaps because there are fewer vehicles to collide with. The paved road has a positive effect on the dependent variable (T = 2.56, p < .0001) while the straight road has a negative effect (T = -17.27, p < .0001).

5 Conclusion

Thanks to open data it is possible to implement useful strategies and share valuable insights about public safety. This study investigates open data about car accidents through the service science lens. Open data initiatives represent a useful tool for public service. It demonstrates the usefulness of a data-oriented culture to increase shareable information assets.

As shown by the analysis, some aspects of the road can incisively impact the outcome of an accident. Mainly the state of the road can be asserted as one of the major points that needs improvement. Furthermore, road signs are not enough to significantly impact the number of victims for road accidents. Finally, our results suggested that the morphology of the carriageways should be reconsidered to diminish the risk of potentially dangerous dynamics such as car skidding.

This paper does not aim at delivering practical indications of best practices for road accidents management. Rather it was designed to explore open data that emerged as a solution in many scenarios such as in the service field to better face several issues being a valid alternative for the future. This article helps at improving our understanding of the usage of open data to define the safety in crowded roads of an important metropolis in the Italian context.

Nevertheless, our study presents some limitations. For instance, the dataset is not representative of the population of accidents in Rome since the recorded accidents exclude the accidents on the "Grande Raccordo Anulare" of Rome and in those where the parties involved have reached a conciliation. Moreover, some important variables were not measured, such as road improvements and the implemented policies (road maintenance, improvements, etc.) during the years under investigation. Further researches could analyse mediation or moderation effects between the variables to include features related to the drivers or to perform a spatial analysis with the accident's location.

6 References

1. J. B. Bullock, "Artificial intelligence, discretion, and bureaucracy," *Am. Rev. Public Adm.*, vol. 49, no. 7, pp. 751–761, 2019.
2. P. A. Busch and H. Z. Henriksen, "Digital discretion: A systematic literature review of ICT and street-level discretion," *Inf. Polity*, vol. 23, no. 1, pp. 3–28, 2018.
3. J. E. Fountain, *Building the virtual state: Information technology and institutional change*. Brookings Institution Press, 2004.
4. E. Barry and F. Bannister, "Barriers to open data release: A view from the top," *Inf. Polity*, vol. 19, no. 1, 2, pp. 129–152, 2014.
5. J. B. Bullock, R. A. Greer, and L. J. O'Toole Jr, "Managing risks in public organizations: A conceptual foundation and research agenda," *Perspect. Public Manag. Gov.*, vol. 2, no. 1, pp. 75–87, 2019.
6. H. Chesbrough and J. Spohrer, "A research manifesto for services science," *Commun. ACM*, vol. 49, no. 7, pp. 35–40, 2006.
7. Ng, R. Maull, and L. Smith, "Embedding the new discipline of service science," in *The science of service systems*, Springer, 2011, pp. 13–35.
8. K. A. Lyons, "Service Science in iSchools," 2010.
9. J. Tidd and F. M. Hull, *Service innovation: Organizational responses to technological opportunities and market imperatives*, vol. 9. World Scientific, 2003.
10. H. De Vries, V. Bekkers, and L. Tummers, "Innovation in the public sector: A systematic review and future research agenda," *Public Adm.*, vol. 94, no. 1, pp. 146–166, 2016.
11. E. M. Rogers, U. E. Medina, M. A. Rivera, and C. J. Wiley, "Complex adaptive systems and the diffusion of innovations," *Innov. J. Public Sect. Innov. J.*, vol. 10, no. 3, pp. 1–26, 2005.
12. S. Borins, "Loose cannons and rule breakers, or enterprising leaders? Some evidence about innovative public managers," *Public Adm. Rev.*, vol. 60, no. 6, pp. 498–507, 2000.
13. L. Brown and S. P. Osborne, "Risk and innovation: Towards a framework for risk

governance in public services," *Public Manag. Rev.*, vol. 15, no. 2, pp. 186–208, 2013.

14. R. L. Ackoff, "From data to wisdom," *J. Appl. Syst. Anal.*, vol. 16, no. 1, pp. 3–9, 1989.
15. J. Rowley, "The wisdom hierarchy: representations of the DIKW hierarchy," *J. Inf. Sci.*, vol. 33, no. 2, pp. 163–180, 2007.
16. M. Jakobsen, O. James, D. Moynihan, and T. Nabatchi, *JPART virtual issue on citizen-state interactions in public administration research*. Oxford University Press US, 2019.
17. T. Nabatchi, "Putting the 'public' back in public values research: Designing participation to identify and respond to values," *Public Adm. Rev.*, vol. 72, no. 5, pp. 699–708, 2012.
18. S. Piotrowski, S. Grimmelikhuijsen, and F. Deat, "Numbers over narratives? How government message strategies affect citizens' attitudes," *Public Perform. Manag. Rev.*, vol. 42, no. 5, pp. 1005–1028, 2019.
19. F. Damanpour, "Organizational innovation: A meta-analysis of effects of determinants and moderators," *Acad. Manage. J.*, vol. 34, no. 3, pp. 555–590, 1991.
20. R. M. Walker, "An empirical evaluation of innovation types and organizational and environmental characteristics: Towards a configuration framework," *J. Public Adm. Res. Theory*, vol. 18, no. 4, pp. 591–615, 2008.
21. R. A. Posner, "From the new institutional economics to organization economics: with applications to corporate governance, government agencies, and legal institutions," *J. Institutional Econ.*, vol. 6, no. 1, p. 1, 2010.
22. J. R. Gil-Garcia, "Towards a smart State? Inter-agency collaboration, information integration, and beyond," *Inf. Polity*, vol. 17, no. 3, 4, pp. 269–280, 2012.
23. Peled, "Coerce, consent, and coax: A review of US congressional efforts to improve Federal Counterterrorism Information Sharing," *Terror. Polit. Violence*, vol. 28, no. 4, pp. 674–691, 2016.
24. J. Lee, "Determinants of government bureaucrats' new PMIS adoption: The role of organizational power, IT capability, administrative role, and attitude," *Am. Rev. Public Adm.*, vol. 38, no. 2, pp. 180–202, 2008.
25. H. Margetts and P. Dunleavy, "The second wave of digital-era governance: a quasi-paradigm for government on the Web," *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.*, vol. 371, no. 1987, p. 20120382, 2013.
26. D. M. West, *Digital government: Technology and public sector performance*. Princeton University Press, 2005.
27. J. Fishenden and M. Thompson, "Digital government, open architecture, and innovation: why public sector IT will never be the same again," *J. Public Adm. Res. Theory*, vol. 23, no. 4, pp. 977–1004, 2013.
28. P. Dunleavy, H. Margetts, S. Bastow, and J. Tinkler, "New public management is dead—long live digital-era governance," *J. Public Adm. Res. Theory*, vol. 16, no. 3, pp. 467–494, 2006.
29. S. G. Grimmelikhuijsen and M. K. Feeney, "Developing and testing an integrative framework for open government adoption in local governments," *Public Adm. Rev.*, vol. 77, no. 4, pp. 579–590, 2017.
30. J. N. Baldwin, R. Gauld, and S. Goldfinch, "What public servants really think of e-government," *Public Manag. Rev.*, vol. 14, no. 1, pp. 105–127, 2012.
31. S. S. Dawes, "Stewardship and usefulness: Policy principles for information-based transparency," *Gov. Inf. Q.*, vol. 27, no. 4, pp. 377–383, 2010.
32. S. Kim and J. Lee, "E-participation, transparency, and trust in local government," *Public Adm. Rev.*, vol. 72, no. 6, pp. 819–828, 2012.
33. J. A. Musso and C. Weare, "Implementing electronic notification in Los Angeles: citizen participation politics by other means," *Int. J. Public Adm.*, vol. 28, no. 7–8, pp. 599–620, 2005.
34. C. J. Tolbert and K. Mossberger, "The effects of e-government on trust and confidence in

government," *Public Adm. Rev.*, vol. 66, no. 3, pp. 354–369, 2006.

35. M. Petychakis, O. Vasileiou, C. Georgis, S. Mouzakitis, and J. Psarras, "A state-of-the-art analysis of the current public data landscape from a functional, semantic and technical perspective," *J. Theor. Appl. Electron. Commer. Res.*, vol. 9, no. 2, pp. 34–47, 2014.
36. J. Thorsby, G. N. Stowers, K. Wolslegel, and E. Tumbuan, "Understanding the content and features of open data portals in American cities," *Gov. Inf. Q.*, vol. 34, no. 1, pp. 53–61, 2017.
37. Zuiderwijk and M. Janssen, "Open data policies, their implementation and impact: A framework for comparison," *Gov. Inf. Q.*, vol. 31, no. 1, pp. 17–29, 2014.
38. T. Nam, "Challenges and concerns of open government: A case of government 3.0 in Korea," *Soc. Sci. Comput. Rev.*, vol. 33, no. 5, pp. 556–570, 2015.
39. Williams, "On the release of information by governments: Causes and consequences," *J. Dev. Econ.*, vol. 89, no. 1, pp. 124–138, 2009.
40. S. Grimmelikhuijsen, "Do transparent government agencies strengthen trust?," *Inf. Polity*, vol. 14, no. 3, pp. 173–186, 2009.
41. S. Wang and M. K. Feeney, "Determinants of information and communication technology adoption in municipalities," *Am. Rev. Public Adm.*, vol. 46, no. 3, pp. 292–313, 2016.
42. L. Wood, P. Bernt, and C. Ting, "Implementing public utility commission web sites: Targeting audiences, missing opportunities," *Public Adm. Rev.*, vol. 69, no. 4, pp. 753–763, 2009.
43. F. Damanpour and M. Schneider, "Characteristics of innovation and innovation adoption in public organizations: Assessing the role of managers," *J. Public Adm. Res. Theory*, vol. 19, no. 3, pp. 495–522, 2009.
44. K. J. Klein and J. S. Sorra, "The challenge of innovation implementation," *Acad. Manage. Rev.*, vol. 21, no. 4, pp. 1055–1080, 1996.
45. H. De Vries, L. Tummers, and V. Bekkers, "The diffusion and adoption of public sector innovations: A meta-synthesis of the literature," *Perspect. Public Manag. Gov.*, vol. 1, no. 3, pp. 159–176, 2018.
46. R. M. Walker, "Internal and external antecedents of process innovation: A review and extension," *Public Manag. Rev.*, vol. 16, no. 1, pp. 21–44, 2014.
47. T. Beshah, D. Ejigu, A. Abraham, P. Krömer, and V. Snášel, "Knowledge Discovery From Road Traffic Accident Data In Ethiopia: Data Quality, Ensembling And Trend Analysis For Improving Road Safety" *Neural Netw. World*, vol. 22, no. 3, pp. 215–244, 2012, doi: 10.14311/NNW.2012.22.013.
48. Y. Zhang, W. Zhou, S. Yuan, and Q. Yuan, "Seizure detection method based on fractal dimension and gradient boosting," *Epilepsy Behav.*, vol. 43, pp. 30–38, Feb. 2015, doi: 10.1016/j.yebeh.2014.11.025.
49. "1.11. Ensemble methods — scikit-learn 0.24.1 documentation." https://scikit-learn.org/stable/modules/ensemble.html#gradient-boosting (accessed Jan. 30, 2021).
50. G. Ciasullo, G. Lodi, A. Maccioni, A. Rotundo, and F. Tortorelli, *Linee Guida Nazionali Per La Valorizzazione Del Patrimonio Informativo Pubblico (anno 2014)*. Technical report, Agenzia per l'Italia Digitale (AgID), Presidenza del …, 2014.
51. Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, 1997.